selecting a Web cache from the plurality of Web caches, the Web cache having service metrics more appropriate for a user request from the Web page than service metrics of other Web caches in plurality of Web caches; and

sending the user request for the static content to the Web cache.

2. The method of Claim 1, wherein the sending step further comprises:

determining traffic loads of a plurality of customer Web servers using a probe server;

selecting the customer Web server from the plurality of customer Web servers using a DNS server, the customer Web server having a traffic load more appropriate for a user request than traffic loads of other customer Web servers in the plurality of customer Web servers; and

sending the user request for the Web page to the customer web server.

3. The method of Claim 2, wherein the traffic loads include latency measurements between the probe server and the plurality of customer Web servers.

4. The method of Claim 2, wherein the determining traffic loads step measures traffic loads at predetermined intervals.

5. The method of Claim 1, wherein the determining service metrics step uses a probe server to determine service metrics of the plurality of Web caches;

wherein the selecting a Web cache step uses a DNS server to select the Web cache from the plurality of Web caches; and

wherein the Web cache sends the static content to the requesting user.

6. The method of Claim 5, wherein the service metrics include metrics selected from: HTTP response time, FTP response time, CPU load, memory load.

7. The method of Claim 5, wherein the determining service metrics step determines service metrics at predetermined intervals.

8. The method of Claim 5, further comprising:

determining whether the requested static content is resident on the Web cache;

determining a customer Web server that has the requested static content when the requested static content is not resident on the Web cache;

wherein the Web cache retrieves the requested static content from the customer Web server; and

storing the requested static content from the customer Web server on the Web cache.

9. The method of Claim 8, wherein the customer Web server from which static content is retrieved is predetermined.

10. The method of Claim 1, wherein the network of POP servers comprises more than one DNS server.

11. A method, comprising:

sending a Web page resident on a customer Web server to a requesting user, the Web page including cacheable content represented by an embedded URL and dynamic content represented by a second embedded URL;

wherein the dynamic content is served by a plurality of customer Web servers;

wherein the cacheable content is served by a plurality of Web caches within a POP server network;

wherein the customer is a customer of a service that operates the plurality of Web caches;

wherein the customer pays a fee to the service for use of the plurality of Web caches storing static content for the customer;

determining service metrics of the plurality of Web caches;

selecting a Web cache from the plurality of Web caches, the Web cache having service metrics more appropriate for a user request from the Web page than service metrics of other Web caches in the plurality of Web caches; and

sending the user request for the static content to the Web cache.

12. The method of Claim 11, wherein the sending step further comprises:

determining traffic loads of the plurality of customer Web servers using a probe server;

selecting the customer Web server from the plurality of customer Web servers using a DNS server, the customer Web server having a traffic load more appropriate for a user request than traffic loads of other customer Web servers in the plurality of customer Web servers; and

sending the user request for the Web page to the customer web server.

13. The method of Claim 12, wherein the traffic loads include latency measurements between the probe server and the plurality of customer Web servers.

14. The method of Claim 12, wherein the determining traffic loads step measures traffic loads at predetermined intervals.

15. The method of Claim 11, wherein the determining service metrics step uses a probe server to determine service metrics of the plurality of Web caches;

wherein the selecting a Web cache step uses a DNS server to select the Web cache from the plurality of Web caches; and

wherein the Web cache sends the static content to the requesting user.

16. The method of Claim 15, wherein the service metrics include metrics selected from: HTTP response time, FTP response time, CPU load, memory load.

17. The method of Claim 15, wherein the determining service metrics step determines service metrics at predetermined intervals.

18. The method of Claim 15, further comprising:

determining whether the requested static content is resident on the Web cache;

determining a customer Web server that has the requested static content when the requested static content is not resident on the Web cache;

wherein the Web cache retrieves the requested static content from the customer Web server; and

storing the requested static content from the customer Web server on the Web cache.

19. The method of Claim 18, wherein the customer Web server from which static content is retrieved is predetermined.

20. The method of Claim 11, wherein the network of POP servers comprises more than one DNS server.

21. (Currently Amended) An apparatus, comprising:

a module for sending a Web page resident on a customer Web server to a requesting user, the Web page including static content represented by an embedded URL;

wherein the static content is served by a plurality of Web caches within a POP server network;

wherein the customer is a customer of a service that operates the plurality of Web caches; ~~and~~

wherein the customer pays a fee to the service for use of the plurality of Web caches storing static content for the customer;

a module for determining service metrics of the plurality of Web caches;

a module for selecting a Web cache from the plurality of Web caches, the Web cache having service metrics more appropriate for a user request from the Web page than service metrics of other Web caches in the plurality of Web caches; and a module for sending the user request for the static content of the Web cache.

22. The apparatus of Claim 21, wherein the sending module further comprises:

a module for determining traffic loads of a plurality of customer Web servers using a probe server;

a module for selecting the customer Web server from the plurality of customer Web servers using a DNS server, the customer Web server having a traffic load more appropriate for a user request than traffic loads of other customer Web servers in the plurality of customer Web servers; and

a module for sending the user request for the Web page to the customer web server.

23. The apparatus of Claim 22, wherein the traffic loads include latency measurements between the probe server and the plurality of customer Web servers.

24. The apparatus of Claim 22, wherein the determining traffic loads module measures traffic loads at predetermined intervals.

25. The apparatus of Claim 21, wherein the module for determining service metrics uses a probe server to determine service metrics of the plurality of Web caches;

wherein the module for selecting a Web cache uses a DNS server to select the Web cache from the plurality of Web caches; and

wherein the Web cache sends the static content to the requesting user.

26. The apparatus of Claim 25, wherein the service metrics include metrics selected from: HTTP response time, FTP response time, CPU load, memory load.

27. The apparatus of Claim 25, wherein the determining service metrics module determines service metrics at predetermined intervals.

28. The apparatus of Claim 25, further comprising:

a module for determining whether the requested static content is resident on the Web cache;

a module for determining a customer Web server that has the requested static content when the requested static content is not resident on the Web cache;

wherein the Web cache retrieves the requested static content from the customer Web server; and

a module for storing the requested static content from the customer Web server on the Web cache.

29. The apparatus of Claim 28, wherein the customer Web server from which static content is retrieved is predetermined.

30. The apparatus of Claim 21, wherein the network of POP servers comprises more than one DNS server.

31. (Currently Amended) An apparatus, comprising:

a module for sending a Web page resident on a customer Web server to a requesting user, the Web page including cacheable content represented by an embedded URL and dynamic content represented by a second embedded URL;

wherein the dynamic content is served by a plurality of customer Web servers;

wherein the cacheable content is served by a plurality of Web caches within a POP server network;

wherein the customer is a customer of a service that operates the plurality of Web caches; ~~and~~

wherein the customer pays a fee to the service for use of the plurality of Web caches storing static content for the customer;

a module for determining service metrics of the plurality of Web caches;

a module for selecting a Web cache from the plurality of Web caches, the Web cache having service metrics more appropriate for a user request from the Web page than service metrics of other Web caches in the plurality of Web caches; and

a module for sending the user request for the static content to the Web cache.

32. The apparatus of Claim 31, wherein the sending module further comprises:

a module for determining traffic loads of the plurality of customer Web servers using a probe server;

a module for selecting the customer Web server from the plurality of customer Web servers using a DNS server, the customer Web server having a traffic load more appropriate for a user request than traffic loads of other customer Web servers in the plurality of customer Web servers; and a module for sending the user request for the Web page to the customer web server.

33. The apparatus of Claim 32, wherein the traffic loads include latency measurements between the probe server and the plurality of customer Web servers.

34. The apparatus of Claim 32, wherein the determining traffic loads module measures traffic loads at predetermined intervals.

35. The apparatus of Claim 31, wherein the module for determining service metrics uses a probe server to determine service metrics of the plurality of Web caches;

wherein the module for selecting a Web cache uses a DNS server to select the Web cache from the plurality of Web caches; and

wherein the Web cache sends the static content to the requesting user.

36. The apparatus of Claim 35, wherein the service metrics include metrics selected from: HTTP response time, FTP response time, CPU load, memory load.

37. The apparatus of Claim 35, wherein the determining service metrics module determines service metrics at predetermined intervals.

38. The apparatus of Claim 35, further comprising:

a module for determining whether the requested static content is resident on the Web cache;

a module for determining a customer Web server that has the requested static content when the requested static content is not resident on the Web cache;

wherein the Web cache retrieves the requested static content from the customer Web server; and

a module for storing the requested static content from the customer Web server on the Web cache.

39. The apparatus of Claim 38, wherein the customer Web server from which static content is retrieved is predetermined.

40. The apparatus of Claim 31, wherein the network of POP servers comprises more than one DNS server.